

# AN ADAPTIVELY WEIGHTED STATISTIC FOR DETECTING DIFFERENTIAL GENE EXPRESSION WHEN COMBINING MULTIPLE TRANSCRIPTOMIC STUDIES

BY JIA LI AND GEORGE C. TSENG<sup>1</sup>

*University of Pittsburgh*

Global expression analyses using microarray technologies are becoming more common in genomic research, therefore, new statistical challenges associated with combining information from multiple studies must be addressed. In this paper we will describe our proposal for an adaptively weighted (AW) statistic to combine multiple genomic studies for detecting differentially expressed genes. We will also present our results from comparisons of our proposed AW statistic to Fisher's equally weighted (EW), Tippett's minimum  $p$ -value (minP) and Pearson's (PR) statistics. Due to the absence of a uniformly powerful test, we used a simplified Gaussian scenario to compare the four methods. Our AW statistic consistently produced the best or near-best power for a range of alternative hypotheses. AW-obtained weights also have the additional advantage of filtering discordant biomarkers and providing natural detected gene categories for further biological investigation. Here we will demonstrate the superior performance of our proposed AW statistic based on a mix of power analyses, simulations and applications using data sets for multi-tissue energy metabolism mouse, multi-lab prostate cancer and lung cancer.

**1. Introduction.** Integrating results from multiple biological studies is now considered commonplace, with significance levels and effect sizes often used in meta-analyses. Random effects models which models effect sizes are frequently used to address variation in sampling schemes. Differences in data structures and statistical hypotheses are common in multiple applications, making direct combinations of effect sizes difficult or impossible. It is more feasible to combine the transformed probability integrals of test statistics (usually  $p$ -values), since the procedure is only dependent

---

Received June 2009; revised July 2010.

<sup>1</sup>Supported in part by NIH (KL2 RR024154-02) and the University of Pittsburgh (Central Research Development Fund, CRDF; Competitive Medical Research Fund, CMRF).

*Key words and phrases.* Meta-analysis, adaptively weighted statistics, genomic study.

This is an electronic reprint of the original article published by the  
 Institute of Mathematical Statistics in *The Annals of Applied Statistics*,  
 2011, Vol. 5, No. 2A, 994–1019. This reprint differs from the original in  
 pagination and typographic detail.

on the significance values of individual tests instead of on underlying data structures. Fisher’s (1932) well-known method of this type involves the log-transformation of  $p$ -values to Chi-square scores and the equally-weighted summation:  $V^{\text{EW}} = -\sum_{k=1}^K \log(p_k)$ , where  $K$  studies are combined and  $p_k$  is the  $p$ -value of study  $k$ ,  $1 \leq k \leq K$ . Assuming independence among studies and  $p$ -values calculated from correct null distributions in each study,  $2V^{\text{EW}}$  follows a Chi-square distribution with  $2K$  degrees of freedom under the null hypothesis. Previously considered other transformations include inverse normal [Stouffer et al. (1949)], logit [Lancaster (1961)] and inverse Chi-square transformation with varying degrees of freedom [George (1977)], among many others. Although Fisher’s method is not the most uniformly powerful, it does exhibit good power for a wide range of conditions. It is also recognized for its asymptotically Bahadur optimal (ABO) characteristic, with multiple studies having the same effect size for alternative hypotheses [Littell and Folks (1971, 1973)]. Different weights or variations of Fisher’s statistic have also been considered. Good (1955) suggested using unequal weights for individual studies in which weights are determined by decisions made by subject experts. More recently, Olkin and Saner (2001) have proposed a trimmed version of Fisher’s statistic to remove the potential effects of aberrant extremes. Another well-known method in the category of combining  $p$ -values is Tippett’s (1931) minimum  $p$ -value statistic (minP):  $V^{\text{minP}} = \min_{1 \leq k \leq K} p_k$ . Wilkinson (1951) generalized Tippett’s procedure to a more robust  $r$ th smallest  $p$ -value, in which  $V^{\text{maxP}} = \max_{1 \leq k \leq K} p_k$  (maxP) is widely used. Note that minP and maxP statistics align with Roy’s (1953) union–intersection test and Berger’s (1982) intersection–union test, respectively. For comprehensive reviews and comparisons of various meta-analysis approaches, see Hedges and Olkin (1985) and Cousins (2007).

Microarray supports the examination of the expression of thousands of genes in parallel. As microarray experiments become more mature and common, it has become increasingly important to integrate homogeneous experimental data sets from multiple laboratories and experimental techniques. In contrast to traditional epidemiological or evidence-based medical studies, the process of monitoring the expression for thousands of genes simultaneously presents many challenges to integrative analysis. In the current biological literature, the term meta-analysis refers to the widespread use of naive intersection/union operations or vote counting on lists of differentially expressed genes obtained from individual studies using certain criteria—for instance, False Discovery Rate  $\leq 0.05$  [Borovecki et al. (2005); Cardoso et al. (2007); Pirooznia, Nagarajan and Deng (2007); Segal et al. (2004), among many others]. Intersections are too conservative and unions insufficiently conservative, especially as the value of  $K$  increases.

More sophisticated meta-analysis methods can be divided into two traditions, the first being the use of a summary statistic—that is, a combination

of statistics from individual studies for each gene being considered, adjusted for multiple comparisons. In many situations, this type of method is an extension of traditional meta-analysis methods. For example, Rhodes et al. (2002), who were the first to apply Fisher’s method to microarray data, later introduced a weighted average of test statistics from individual tests, with weights determined by study sample sizes [Ghosh et al. (2003)]. Moreau et al. (2003) made use of Tippet’s minimum  $p$ -value. A more robust statistic is Wilkinson’s  $r$ th smallest  $p$ -value, in which maximum  $p$ -value can be applied to the meta-analysis of microarray studies. Owen (2009) reintroduced Pearson’s (1934) method and applied it to the AGEMAP project. He defined a test statistic as the maximum of Fisher’s combination of left-sided and right-sided  $p$ -values. All of these methods combine statistical significance. Note that when no gene effect exists, the  $p$ -value is uniformly distributed. Accordingly, combining the significance of independent tests is sometimes called omnibus or nonparametric. When studies have similar design and measure the outcomes in similar ways, combining effect sizes is usually preferred to combining significance. Choi et al. (2003) used weighted estimate for individual genes based on the random effects model (REM) under Gaussian assumptions, and discussed the details of a Bayesian formulation for the REM model. Hu, Greenwood and Beyene (2005) developed a quality measure for each gene in individual studies, incorporating a quality index as a weight in the REM model. Hong et al. (2006) proposed a robust rank-based approach for meta-analysis. Choi et al. (2007) introduced a latent variable approach.

The second meta-analysis tradition is Bayesian—for example, Choi et al.’s (2003) Bayesian version for REM, which models the effect sizes. Similar Bayesian hierarchical models have been suggested by Tseng et al. (2001) and Conlon, Song and Liu (2006) for incorporating different levels of replicates information in cDNA microarray experiments. Conlon, Song and Liu (2007) refer to these models as Bayesian probability integration (PI) models, and have introduced a Bayesian standardized expression integration (SEI) model. Instead of modeling study specific means separately (PI model), SEI models them as samples from a normal distribution, thus producing overall mean and inter-study variation. Shen, Ghosh and Chinnaiyan (2004) and Choi et al. (2007) used a Bayesian mixture model to rescale the individual data set and then combined all data sets for an ordinary gene expression analysis.

The structure for the rest of this paper is as follows: in Section 2 we describe two complementary hypothesis settings for detecting study-invariant and study-specific biomarkers:  $HS_A$  and  $HS_B$ . In Section 3 we present our proposal for an adaptively weighted (AW) statistic for meta-analyses of genomic studies, including detailed descriptions of the AW statistic algorithm and a permutation test for combining multiple studies. In Section 4 we discuss a simulation test of our proposed method, using data sets from studies

of a multi-tissue energy metabolism mouse model, prostate cancer and lung cancer; we then compare our results with those produced by three other commonly used methods. In Section 5 we demonstrate the admissibility and power of our proposed AW test under a Gaussian assumption, and in Section 6 we summarize its statistical advantages and limitations.

**2. Two major complementary hypothesis settings.** To our knowledge, no comprehensive evaluations for the above-described meta-analysis methods have been performed, primarily due to a lack of rigorous formulation of statistical hypotheses. Here we will consider a meta-analysis of  $D_1, D_2, \dots, D_K$  gene expression profiles studies.  $x_{kgs}$  is the gene expression intensity of gene  $g$  and sample  $s$  in study  $k$ , with samples  $s = 1, \dots, n_k$  belonging to a control group (e.g., normal samples) and  $s = n_k + 1, \dots, n_k + m_k$  belonging to the diseased group (e.g., cancer samples). Normally a null hypothesis for each gene  $g$  is considered as

$$H_0: \theta_{g1} = \dots = \theta_{gK} = 0,$$

where  $\theta_{gk}$  represents the gene effect of gene  $g$  and study  $k$ . Building on Birnbaum's (1954) work, the complementary hypothesis settings ( $HS_A$  and  $HS_B$ ) are dependent upon the nature of the experiment in which the gene effects ( $\theta_{gk}$ ) are obtained:

$$HS_A: \{H_0 \text{ versus } H_A: \theta_{gk} \neq 0, \forall 1 \leq k \leq K\},$$

$$HS_B: \{H_0 \text{ versus } H_B: \text{at least one } \theta_{gk} \neq 0, 1 \leq k \leq K\}.$$

It is possible to use different methods to explicitly or implicitly consider different subsets or variations of the two alternative hypotheses:

$$HS_{A1}: \{H_0 \text{ versus } H_{A1}: \theta_g = \theta_{g1} = \dots = \theta_{gK} \neq 0\},$$

$$HS_{A2}: \{H_0 \text{ versus } H_{A2}: \theta_g \neq 0, \theta_{gk} \sim N(\theta_g, \tau^2)\},$$

$$HS_{Bh}: \{H_0 \text{ versus } H_{Bh}: \sum_{k=1}^K I(\theta_{gk} \neq 0) = h \ (1 \leq h \leq K)\}$$

[ $I(\cdot)$  is an indicator function that

equals 1 when statement true and 0 otherwise],

$$HS_{Bh'}: \left\{ H_0 \text{ versus } H_{Bh'}: \sum_{k=1}^K I(\theta_{gk} \neq 0) = h \right. \\ \left. \text{and } \theta_{gk} = \theta_g \text{ if } \theta_{gk} \neq 0 \ (1 \leq h \leq K) \right\}.$$

Without danger of confusion, here we will use  $H_A$  notation to denote the parameter space of the corresponding alternative hypothesis. It is clearly seen that  $H_A \subset H_B$ . However, they represent two families of complementary interpretations in applications. Under  $H_A$ , gene  $g$  is identified only when it is differentially expressed in all studies. Under  $H_B$ , gene  $g$  is selected only if it is differentially expressed in one or more studies. Note that  $H_{A1} \subset H_A$ , representing an equal fixed effect model.  $H_{A2}$  represents a random effects model for a similar  $H_A$  purpose, while  $H_{A2} \not\subset H_A$  in general. Note also that  $H_B = \bigcup_{1 \leq h \leq K} H_{Bh}$ ,  $H_{Bh'} \subset H_{Bh}$  ( $1 \leq h \leq K$ ) and  $H_{BK'} = H_{A1}$ .

From a biological standpoint, experimental design and meta-analysis objectives determine biomarker lists of interest. To illustrate this idea, we will use three sets of microarray studies for meta-analyses. The first set consists of two mouse genotypes, wild type (VLCAD +/+) and VLCAD deficient (VLCAD -/-), with four mice in each genotype group (VLCAD is associated with a childhood metabolism disorder). Brown fat, liver and heart tissue samples were collected from each of the eight individual mice and used for microarray experiments designed to study global expression changes in the knock-out of VLCAD (Table 1, left). Given the experimental design, a biomarker list of interest might consist of those genes that are consistently expressed in all tissue samples from both wild type and VLCAD-deficient mice. This type of tissue-invariant (or study-invariant) biomarker list can be loosely defined as  $G_A$ , with analysis based on the alternative hypothesis family of  $H_A$ . However, it is reasonable to assume that tissue-specific physiology triggers tissue-dependent responses, with pools of differentially expressed genes being confounded to the tissues in question. Such a hypothesis would focus on signature genes that are differentially expressed in subsets of one or more tissues—an analysis that corresponds to the  $H_B$  alternative hypothesis family. Hereafter we will use the term  $G_B$  when addressing such tissue-specific or study-specific biomarker lists. In the second study set, microarray comparisons of normal versus prostate tumor tissues were performed by three different research teams: Dhanasekaran et al. (2001), Luo et al. (2001) and Welsh et al. (2001) (Table 1). The  $G_A$  study-invariant biomarker list is clearly of greater biological interest in this situation, since many of the  $G_B$  study-specific biomarkers represent experimental and technical discrepancies between studies, possibly due to sample population heterogeneity, gene matching errors or differences in experimental protocols. Further investigation of study-specific biomarkers may provide technical insights to experimental design features without providing biological insights to the disease of interest. The third set of microarray studies [Bhattacharjee et al. (2001); Beer et al. (2002); Garber et al. (2001)] included analyses of lung cancer samples and a comparison of normal versus adenocarcinoma samples. Table 1C shows the pair-wise integrative correlation coefficients [Parmigiani et al. (2004)] in each of the three examples. A review of past

TABLE 1  
Three sets of microarray studies for meta-analyses. (BF—brown fat; Liv—liver; Ht—heart; WT—wild type (VLCAD +/+); VLCAD—VLCAD −/−; N—normal; T—tumor; AC—adenocarcinomas)

(A)	Mouse energy metabolism			Prostate cancer studies				Lung cancer studies				
		BF	Liv	Ht		Dhan	Luo	Wels		Bhat	Beer	Garb
	WT	4	4	3	N	19	9	9	N	17	10	5
	VLCAD	4	4	4	T	14	16	25	AC	134	86	39
(B)												
$HS_A$		Of biological interest				Of biological interest				Of biological interest		
$HS_B$		Of biological interest				Of less biological interest but of more technical interest				Of less biological interest but of more technical interest		
(C)												
	BF	1	0.06	0.04	Dhan	1	0.05	0.09	Bhat	1	0.33	0.22
	Liv	0.06	1	0.03	Luo	0.05	1	0.09	Beer	0.33	1	0.15
	Ht	0.04	0.03	1	Wels	0.09	0.09	1	Garb	0.22	0.15	1

TABLE 2  
*Meta-analysis methods, corresponding hypothesis settings and targeted types of biomarker list*

Methods	Abbreviation	Alternative hypothesis	Targeted biomarker list
Fisher [equally weighted sum of $\log(p\text{-values})$ ]	EW	$H_B$	$G_B$
Tippett (minimum $p\text{-value}$ )	minP	$H_B$	$G_B$
Pearson (maximum of Fisher's left-sided and right-sided score)	PR	$H_B$	$G_B$
Li and Tseng [adaptively weighted sum of $\log(p\text{-values})$ ]	AW	$H_B$	$G_B$
Wilkinson (maximum $p\text{-value}$ )	maxP	$H_A$	$G_A$
Choi (2003); Shen (2004); Choi (2007) (random effects model)	REM	$H_{A2}$	$G_A$
Conlon (2006) (PI Bayesian approach)	PI	NA	$G_A$
Conlon (2007) (SEI Bayesian approach)	SEI	NA	$G_A$

meta-analyses reveals that lung cancer studies generally have larger samples, greater homogeneity and better data quality than prostate cancer studies, especially in terms of biomarker detection and classification analysis.

Table 2 presents a list of commonly used meta-analysis methods for microarray studies, their corresponding alternative hypotheses and targeted biomarkers. While both Bayesian SEI and PI methods tend to detect  $G_A$ -type biomarkers across studies, the Bayesian concept does not involve hypothesis testing. Note that different approaches have distinctly different advantages and disadvantages in terms of parameter space subsets in alternative hypotheses, even though two methods may be designed for the same hypothesis. For example, to detect  $G_A$  genes, PI performs better than SEI for genes that have a high mean effect in one study but low mean effect in another. According to Laughin (2004), maxP is generally under-powered, but performs well when all  $\theta_{gk}$  values are nonzero and roughly the same. As we will show in Section 5, EW, minP, PR and AW are all admissible for detecting  $G_B$  genes. For  $H_{Bh}$ , EW tends to be more powerful when  $h$  is large and closer to  $K$ . Little and Folks proved that EW is asymptotically ABO when detecting  $G_A$ -type genes under under  $H_{BK'}$  (i.e.,  $H_{A1}$ ), even though the EW statistic is targeted toward general  $H_B$ . In contrast, minP is more powerful in detecting genes under  $H_{Bh}$  when  $h$  is small.

From this point forward, our focus will be on the  $H_B$  alternative hypothesis. In the following section we will describe our proposal for an adaptively weighted statistic (AW), and, in Section 5, we will demonstrate its robustness and near-optimal power for alternative hypotheses at either extreme

(i.e., when  $h$  is close to  $K$  or close to 1 in  $H_{Bh'}$ ). We will also give examples of situations in which AW outperforms EW and minP in intermediate scenarios. AW is capable of distinguishing  $G_A$  and  $G_B \setminus G_A$  genes in a manner that indicates in which study or studies individual biomarkers are differentially expressed—information considered useful for post-meta-analysis investigations.

**3. Adaptively-weighted statistic.** When integrating multiple genomic studies, expression of some important biomarkers may be altered in a study-specific manner (consider  $H_B$ ). To uncover altered gene expression patterns across studies, we start with the following weighted statistic:

$$(3.1) \quad U_g(w_g) = - \sum_{k=1}^K w_{gk} \log(p_{gk}),$$

where  $p_{gk}$  is the  $p$ -value of gene  $g$  in study  $k$ ,  $w_k$  is the weight assigned to the  $k$ th study and  $w_g = (w_{g1}, \dots, w_{gK})$ . Under the null hypothesis that  $\theta_{gk} = 0 \ \forall k$ , the  $p$ -value of the observed weighted statistic,  $p_U(u_g(w_g))$ , can be obtained for a given gene  $g$  and weight  $w_g$  (see below for detailed permutation algorithm to calculate the  $p$ -value). The adaptively-weighted statistic is defined as the minimal  $p$ -value among all possible weights:

$$(3.2) \quad V_g^{\text{AW}} = \min_{w_g \in W} p_U(u_g(w_g)),$$

where  $u_g(w)$  is the observed statistic for  $U_g(w)$ , and  $W$  is a prespecified search space. Our choice of search space in this paper is  $W = \{w \mid w_i \in \{0, 1\}\}$ , which results in an affordable computation of  $O(2^K - 1)$  based on the norm of  $K \leq 10$  in a microarray meta-analysis.

The resulting weight reflects a natural biological interpretation of whether or not a study contributes to the statistical significance of a gene. Note that the AW statistic is inadequate for traditional meta-analysis in epidemiological or evidence-based medicine research. The AW selection procedure will introduce selection bias toward studies with concordant significant effects. However, integrative analysis of genomic studies represents a different situation: usually the primary goal is to screen and identify the most probable gene markers, given data meant to facilitate future investigation. As we will show in Section 4, the weight vector,  $w_g^* = \arg \min_{w_g \in W} p_U(u_g(w_g))$ , actually serves as a convenient basis for gene categorization in follow-up biological interpretations and explorations.

Below we illustrate the detailed procedure for AW when applied to combined genomic studies. If assuming  $p_{gk} \sim \text{Unif}[0, 1]$  under the null hypothesis,  $U_g(w_g) \sim \text{Gamma}(\sum_{k=1}^K w_{gk}, 1)$  and inference of the AW statistic can be performed on this basis. Such a uniform  $p$ -value assumption is, however,



usually not true in real applications. Alternatively, a permutation test is performed below to assess the statistical significance and the false discovery rate (FDR) is controlled at 5%. For the applications in Section 4, the EW, minP, maxP and PR methods are performed using a similar permutation test.

I. Study-wise  $p$ -value calculation before meta-analysis:

- (1) Compute the penalized  $t$ -statistics,  $t_{gk}$ , for gene  $g$  and study  $k$  [Efron et al. (2001); Tusher, Tibshirani and Chu (2001)].
- (2) Permute group labels in each study for  $B$  times, and similarly calculate the permuted statistics,  $t_{gk}^{(b)}$ , where  $1 \leq g \leq G, 1 \leq k \leq K, 1 \leq b \leq B$ .
- (3) Estimate the  $p$ -value of  $t_{gk}$  as  $p_{gk} = (\sum_{b=1}^B \sum_{g'=1}^G I(t_{g'k}^{(b)} \in R(t_{gk}))) / (B \cdot G)$ , where  $R(t_{gk})$  is the rejection region given the threshold  $t_{gk}$ . Similarly, given  $t_{gk}^{(b)}$ , compute  $p_{gk}^{(b)} = (\sum_{b'=1}^B \sum_{g'=1}^G I(t_{g'k}^{(b')} \in R(t_{gk}^{(b)}))) / (B \cdot G)$ .

II. Calculate AW statistic:

- (1) Given a weight  $w_g = (w_{g1}, \dots, w_{gK})$ , the weighted statistic is defined as  $u_g(w_g) = -\sum_{k=1}^K w_{gk} \log(p_{gk})$  for gene  $g$ . Define  $u_g^{(b)}(w_g) = -\sum_{k=1}^K w_{gk} \log(p_{gk}^{(b)})$ .
- (2) Estimate the  $p$ -value of the observed  $u_g(w_g)$  as

$$p_U(u_g(w_g)) = \frac{\sum_{b=1}^B \sum_{g'=1}^G I\{u_{g'}^{(b)}(w_g) \geq u_g(w_g)\}}{B \cdot G}.$$

Similarly compute

$$p_U(u_g^{(b)}(w_g)) = \frac{\sum_{b'=1}^B \sum_{g'=1}^G I\{u_{g'}^{(b')}(w_g) \geq u_g^{(b)}(w_g)\}}{B \cdot G}.$$

- (3) Based on II(1) and II(2), calculate the optimal weight as

$$w_g^* = \arg \min_{w_g \in W} p_U(u_g(w_g))$$

and, similarly,

$$w_g^{(b)*} = \arg \min_{w_g \in W} p_U(u_g^{(b)}(w_g)).$$

Define the AW statistic  $V_g$  as the  $p$ -value of the adaptively weighted statistic:  $V_g = p_U(u_g(w_g^*))$ . Similarly,  $V_g^{(b)} = p_U(u_g^{(b)}(w_g^{(b)*}))$ .

III. Assess  $p$ -values and  $q$ -values of the AW statistic— $V_g$ :

- (1) The  $p$ -value of  $V_g$  is calculated as

$$p_V(V_g) = \frac{\sum_{b=1}^B \sum_{g'=1}^G I\{V_{g'}^{(b)} \leq V_g\}}{B \cdot G}.$$

- (2) Estimate  $\pi_0$ , the proportion of null genes, as

$$\hat{\pi}_0 = \frac{\sum_{g=1}^G I\{p_V(V_g) \in A\}}{G \cdot \ell(A)}$$

[Storey (2002)]. Normally we choose  $A = [0.5, 1]$  and  $\ell(A) = 0.5$ .

- (3) Estimate the  $q$ -value for each gene as

$$q(V_g) = \frac{\hat{\pi}_0 \sum_{b=1}^B \sum_{g'=1}^G I\{V_{g'}^{(b)} \leq V_g\}}{B \sum_{g'=1}^G I\{V_{g'} \leq V_g\}}.$$

The detected gene list is  $G^{\text{AW}} = \{g : q_V(V_g) \leq 0.05\}$ .

- IV. Distinguish concordant and discordant genes (recommended): Split the detected gene list  $G^{\text{AW}}$  into concordant and discordant gene lists. By controlling the false discovery rate (FDR) at 5%, detected genes with concordant regulation direction across contributing studies are denoted as  $G_{\text{concordant}}^{\text{AW}} = \{g : q(V_g) \leq 0.05 \text{ and } |\sum_{k=1}^K \text{sgn}(t_{gk}) \cdot w_{gk}^*| = \sum_{k=1}^K w_{gk}^*\}$ , where  $\text{sgn}(\cdot)$  is the sign function that takes value 1 when positive and  $-1$  when negative. The discordant gene list is  $G_{\text{discordant}}^{\text{AW}} = G^{\text{AW}} \setminus G_{\text{concordant}}^{\text{AW}}$ .

#### REMARKS.

1. For the application of EW and the minP, maxP and PR method, steps II(1)–II(3) can be skipped. Alternatively, the test statistics are modified as  $V_g = -\sum_{k=1}^K \log(p_{gk})$  for EW;  $V_g = \min_{1 \leq k \leq K} p_{gk}$  for minP;  $V_g = \max_{1 \leq k \leq K} p_{gk}$  for maxP and  $V_g = \max(-\sum_{k=1}^K \log(\tilde{p}_{gk}), -\sum_{k=1}^K \log(1 - \tilde{p}_{gk}))$  for PR, where  $\tilde{p}_{gk}$  is the one-sided  $p$ -value for gene  $g$  in study  $k$ .
2. The I–III sequence provides an algorithm for a general framework. Both statistics  $t_{gk}$  and rejection region  $R(t_{gk})$  can be replaced, depending on the experimental design and hypothesis. For example, the  $F$ -statistic can be used when multiple groups of samples are available in each study under consideration.
3. When conducting comparisons of two groups and applying the moderated  $t$ -statistic, genes detected under the general framework (the I–III sequence) may contain discordant genes—for instance, a gene up-regulated in one study and down-regulated in another; the addition of step IV provides further filtering. In some applications, a researcher may want to scrutinize the discordant gene list to verify whether the discordance reflects actual biological discrepancy across studies (e.g., different tissues or patient populations) or artificial errors (e.g., mistakes in gene annotation). For EW and minP there is no direct criterion for a clear split of concordant and discordant genes. After revisiting the PR method for the AGEMAP project, Owen found that it is sensitive to consistent left- or right-sided departures. The PR method is still easily dominated by one or

two exceptionally significant  $p$ -values, and does not identify which studies are significant in distinguishing between concordant versus discordant patterns (see first two examples in Table 6).

4. Several forms of penalized or moderated  $t$ -statistics have been proposed and shown to outperform traditional  $t$ -statistics [Efron et al. (2001); Tusher, Tibshirani and Chu (2001); Smyth (2004)]. For our algorithm we choose the penalized  $t$ -statistics used in Efron et al. (2001) and Tusher, Tibshirani and Chu (2001). The fudge parameter  $s_0$  is chosen to be the median variability estimator in the genome.

#### 4. Applications.

4.1. *Simulation study.* We conducted a simulation study for combining four data sets to compare the performance among our proposed AW test, Fisher’s EW test, Tippett’s minP method, Wilkinson’s maxP method and Pearson’s statistic (PR). For each data set, we simulated five normal samples from a standard normal distribution and five case samples from  $N(\theta, 1)$ . A total of  $g_1$  genes (category I) were differentially expressed across all four data sets;  $g_2 = 400 - g_1$  genes were differentially expressed in the fourth data set only (category II); and 1600 genes were considered null. Genes are called significant by controlling FDR at 5% for each method. Each simulation scenario was repeated 1000 times.

Summaries of the resulting FDR and average number of genes identified in each category under three different scenarios appear in the following tables: 0 category I and 400 category II genes in Table 4; 200 category I and 200 category II genes in Table 5; 400 category I and 0 category II genes in Table 3. The results are consistent with the power calculation discussed in Section 5.1. In Table 3, minP is much more powerful than EW. When  $\theta = 2$ , minP correctly detects an average of 41.6 genes and EW detects only

TABLE 3  
*Evaluation of AW, EW, minP, maxP and PR methods by simulations in the first scenario (I. 0 common DE genes; II. 400 4th-data set-specific DE genes; Null. 1600 random noise genes). Average number of genes detected in each category and the average FDR are shown under different effect size  $\theta$*

Methods	$\theta = 2.0$				$\theta = 2.5$			
	I	II	Null	FDR (s.e.)	I	II	Null	FDR (s.e.)
AW	0.0	32.1	1.9	4.8% (0.002)	0.0	137.1	7.5	4.9% (0.001)
EW	0.0	7.6	0.4	4.1% (0.003)	0.0	43.1	2.0	4.2% (0.002)
minP	0.0	41.6	2.4	5.0% (0.002)	0.0	163.0	8.7	4.9% (0.001)
maxP	0.0	0.2	0.1	25.5% (0.013)	0.0	0.2	0.1	25.5% (0.013)
PR	0.0	3.2	0.1	3.7% (0.004)	0.0	15.2	0.4	2.2% (0.002)

TABLE 4

*Evaluation of AW, EW, minP, maxP and PR methods by simulations in the second scenario (I. 200 common DE genes; II. 200 4th-data set-specific DE genes; Null. 1600 random noise genes). Average number of genes detected in each category and the average FDR are shown under different effect size  $\theta$*

Methods	$\theta = 1.5$				$\theta = 2.0$			
	I	II	Null	FDR (s.e.)	I	II	Null	FDR (s.e.)
AW	169.1	24.3	10.1	4.9% (0.0005)	198.7	59.4	13.4	4.9% (0.0004)
EW	188.4	16.9	8.5	4.0% (0.0004)	199.8	35.4	9.5	3.9% (0.0004)
minP	25.4	6.9	1.9	5.0% (0.0016)	144.0	54.7	10.3	4.9% (0.0005)
maxP	168.3	3.7	8.4	4.6% (0.0005)	195.7	4.4	9.8	4.7% (0.0005)
PR	178.7	9.4	3.8	2.0% (0.0003)	199.3	21.3	4.3	1.9% (0.0003)

7.6 genes. AW detects 32.1 genes, considerably close to minP. Similarly, in Table 5, EW (386.8 genes are detected when  $\theta = 1.5$ ) is more powerful than minP (121.3 genes detected) and AW (359.3 genes detected) is close to EW in performance. Overall, AW performance was stable in these extreme situations. We note most methods show FDR close to 5%, although maxP loses so much power at scenario 1 that FDR is inflated and the PR method appears slightly conservative.

**4.2. Energy metabolism in mouse model.** An energy metabolism disorder in children is associated with very longchain acyl-coenzyme A dehydrogenase (VLCAD) deficiencies. In an ongoing unpublished project, two genotypes of the mouse model—wild type (VLCAD +/+) and VLCAD-deficient (VLCAD -/-)—were studied for three types of tissues (brown fat, liver and heart) with 4 mice in each genotype group. Microarray experiments were applied separately to study the expression changes across genotypes.

TABLE 5

*Evaluation of AW, EW, minP, maxP and PR methods by simulations in the third scenario (I. 400 common DE genes; II. 0 4th-data set-specific DE genes; Null. 1600 random noise genes). Average number of genes detected in each category and the average FDR are shown under different effect size  $\theta$*

Methods	$\theta = 1.5$				$\theta = 2.0$			
	I	II	Null	FDR (s.e.)	I	II	Null	FDR (s.e.)
AW	359.3	0.0	18.6	4.9% (0.0004)	398.5	0.0	20.4	4.8% (0.0004)
EW	386.8	0.0	15.9	4.0% (0.0003)	399.8	0.0	16.1	3.9% (0.0003)
minP	121.3	0.0	6.3	4.8% (0.0007)	329.5	0.0	16.8	4.8% (0.0004)
maxP	357.5	0.0	19.0	5.0% (0.0004)	394.9	0.0	21.3	5.1% (0.0004)
PR	373.9	0.0	7.5	2.0% (0.0002)	399.4	0.0	7.8	1.9% (0.0002)

TABLE 6

Five genes from the mouse energy metabolism data. Moderated  $t$ -statistics and  $p$ -values for individual studies are listed.  $w*$  represents AW-obtained weight. AW2 represents AW concordant method

Gene	Moderated $t$ -statistic ( $p$ -value)			Is it detected ( $q(V) \leq 5\%$ )?					Concordant?
	Brown fat	Liver	Heart	EW	minP	PR	AW	AW2	
1423407_a_at	2.2 (0.0027)	1.7 (0.0027)	-3.7 (0.0014)	✓	×	✓	✓	×	no
$w*$	1	1	1						
1418429_at	3.6 (0.0003)	1.1 (0.067)	-3.2 (0.002)	✓	×	✓	✓	×	no
$w*$	1	0	1						
1449015_at	0.4 (0.46)	-3.3 (0.0009)	-1.8 (0.011)	✓	×	✓	✓	✓	yes
$w*$	0	1	1						
1416415_a_at	-0.8 (0.15)	2.2 (0.0026)	2.6 (0.0023)	✓	×	✓	✓	✓	yes
$w*$	0	1	1						
1415727_at	-1.5 (0.018)	-1.6 (0.014)	-3.5 (0.0008)	✓	×	✓	✓	✓	yes
$w*$	1	1	1						

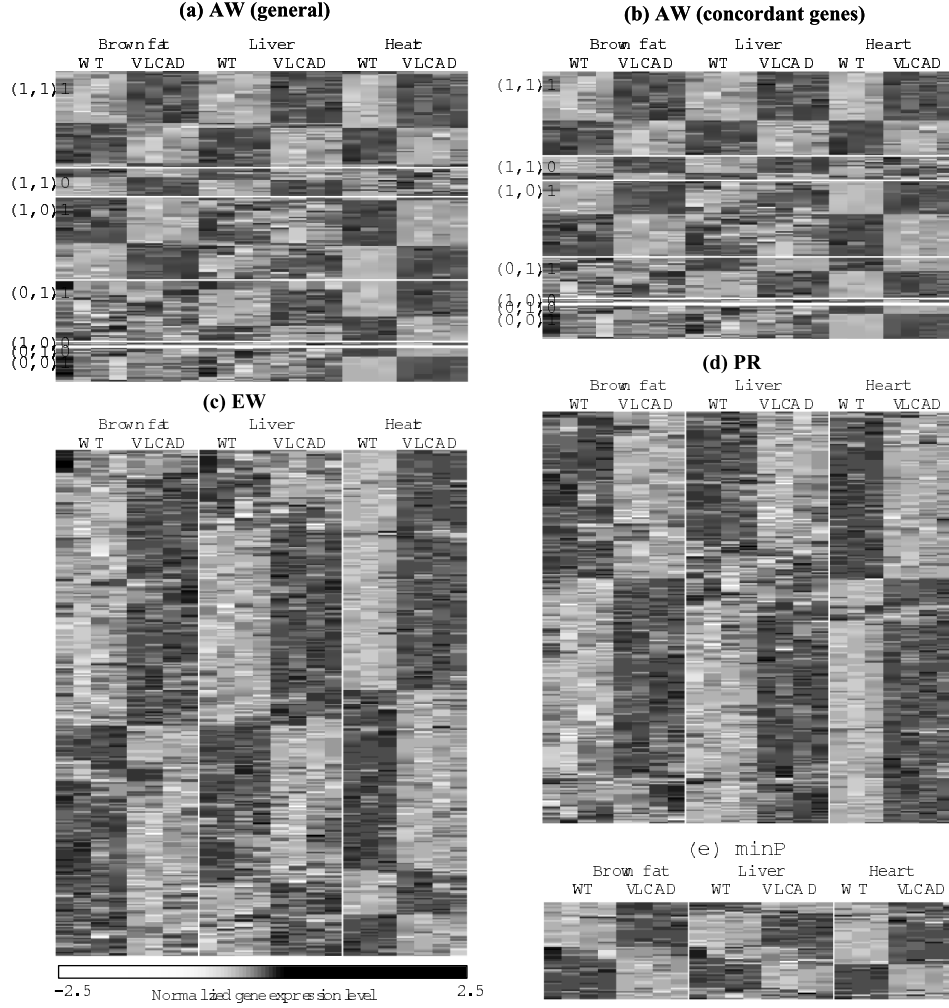


FIG. 1. Heatmaps of gene expressions for differentially expressed genes identified by different methods in the mouse energy metabolism data sets.

In this study we tested the hypotheses that tissue-specific physiology triggers tissue-dependent responses, with precise pools of differentially expressed genes specific to the tissue in question. The purpose of this hypothesis is to identify signature genes that are significant for tissue subsets—an analysis that corresponds to  $HS_B$ .

Due to the low power of maxP, the Figure 1 data are limited to AW, EW, minP and PR methods. Note that EW, minP and AW are based on the summarization of  $p$ -values across studies, and that the methods alone do not distinguish among discordant genes with difference in expression across studies

(e.g., up-regulated in one study but down-regulated in another). The modified algorithm of AW for filtering out discordant genes (Section 3, step IV) can be implemented in such situations, since it discards all discordant genes among studies that contribute to the adaptive weight. The modified AW algorithm is not applicable to EW, minP and PR because those methods do not provide which studies should be considered for concordance/discordance evaluations.

Overall, the general AW detects 203 genes [Figure 1(a)]; among these, 28 genes were conflicting in terms of up- or down-regulations—for example, Figure 1(b) shows the detection of 175 genes. Adaptive-weights serve as a natural grouping process for identified genes: 55 genes with weights of  $(1, 1, 1)$  are differentially expressed in all three tissue types [Figure 1(b)], and 27 with weights of  $(0, 1, 1)$  were differentially expressed in liver and heart tissues, but not in brown fat. The number of detected genes related to heart tissue [ $(1, 1, 1)$ ,  $(1, 0, 1)$ ,  $(0, 1, 1)$  and  $(0, 0, 1)$  in Figure 1(b)] is much higher than that related to brown fat or liver tissues, representing increase impact of VLCAD deletion in heart metabolism activities. According to the EW results shown in Figure 1(c), that method detected more genes (329) than our proposed AW method. However, the identified gene list is difficult to interpret and investigate, even after reordering by hierarchical clustering. In this application minP appears to be much less powerful.

To illustrate AW performance in terms of genes that consistently regulate in the same direction across data sets, details for five genes are presented in Table 6. Four of the five methods identified the five example genes as differentially expressed (the exception was minP). The first two genes (1423407\_a\_at and 1418429\_at) clearly indicate discordant regulation with opposite moderated  $t$ -statistics between brown fat and heart. Even though Pearson’s method (PR) was specifically designed to detect concordant genes, it failed to achieve this goal in this particular situation. In contrast, our proposed AW method uses a post-hoc approach (Section 3, step IV) to filter out discordant genes. Such a post-hoc procedure is not feasible for EW, minP or PR without indicating which studies are differentially expressed. For example, in 1449015\_at and 1416415\_a\_at, the AW method with concordance filtering will still identify them as concordant DE genes, even though regulation of the nonsignificant study (brown fat) contradicts the two significant studies. The difference between AW and the natural tendency of biologists to pick studies based on  $p$ -values obtained from individual analysis is illustrated by the fifth gene, 1415727\_at, which produces moderate signals for brown fat and liver tissue and a very strong signal for heart tissue, to the degree that it can easily be ignored for brown fat and liver following adjustment for multiple comparisons. It is, in general, difficult to decide whether it is a  $(0, 0, 1)$ - or  $(1, 1, 1)$ -type of gene. The fact that this gene is moderately significant in two studies and very significant in a third study enabled AW

to determine that combining results across all three studies gives the best statistical significance and it should be a  $(1, 1, 1)$ -type of gene.

*4.3. Prostate cancer and lung cancer studies.* We applied the AW, EW, minP and PR methods to three sets of prostate cancer data and three sets of lung cancer data (Table 1). Some of the studies were performed by cDNA technology [Dhanasekaran et al. (2001), Luo et al. (2001) and Garber et al. (2001)] while others used Affymetrix oligo-based technology [Welsh et al. (2001), Bhattacharjee et al. (2001) and Beer et al. (2002)]. Data set probes were matched according to their Entrez IDs; the intensities of multiple probes matching the same ID were averaged. For the prostate cancer data set, comparisons were made between clinically localized cancer and benign tissues. For the lung cancer data set we compared tissues from adenocarcinoma patients with those from healthy donors.

The results shown in Figures 2 and 3 reflect characteristics that are similar to those discussed in the above mouse example. With an exception, minP did not perform as poorly as it did in Section 4.1. Compared to the other methods, our proposed AW method identified much clearer patterns. Of the 722 genes in Figure 2(a), 618 genes show consistent regulation across studies [Figure 2(b)]. Approximately 14% of the identified genes were discordant across studies. Possible causes of discordant genes may include mistaken gene annotations in old array platforms [Dai et al. (2005)], differential probe efficiencies, heterogeneous sample populations across studies and nonspecific cross hybridizations. According to our findings, only moderately concordant information existed across the three prostate cancer studies, probably because (a) their sample sizes were small, or (b) they entailed in-house cDNA arrays or commercial products that were still in the early stages of development. Of the 618 concordant AW-detected genes, 130 genes (21%) were consistent  $(1, 1, 1)$ -type biomarkers and 205 genes (33.2%) were specific to one study only: 55  $(1, 0, 0)$ -type biomarkers, 70 of the  $(0, 1, 0)$  type, and 80 of the  $(0, 0, 1)$  type. The EW, minP and PR methods all detected slightly greater numbers of biomarkers than the AW method (924, 745 and 882, resp.). However, in each case the detected biomarkers were difficult to interpret and follow up, and all three methods presented challenges in terms of guaranteeing the detection of concordant genes only. In summary, our findings suggest that results from individual microarray studies require careful interpretation, and that integrative analyses are appropriate as a validation tool.

Similar patterns and results were obtained when the four methods were applied to lung cancer studies (Figure 3). The AW method detected 366 genes, with 349 confirmed as concordant (only 4.6% are discordant compared to 14.4% in prostate cancer). Among the 349 concordant biomarkers, 99 were type  $(1, 1, 1)$  (28.4% compared to 21% in prostate cancer) and 96 were single study specific (27.5% compared to 33.2% in prostate cancer): 7 type  $(1, 0, 0)$ ,



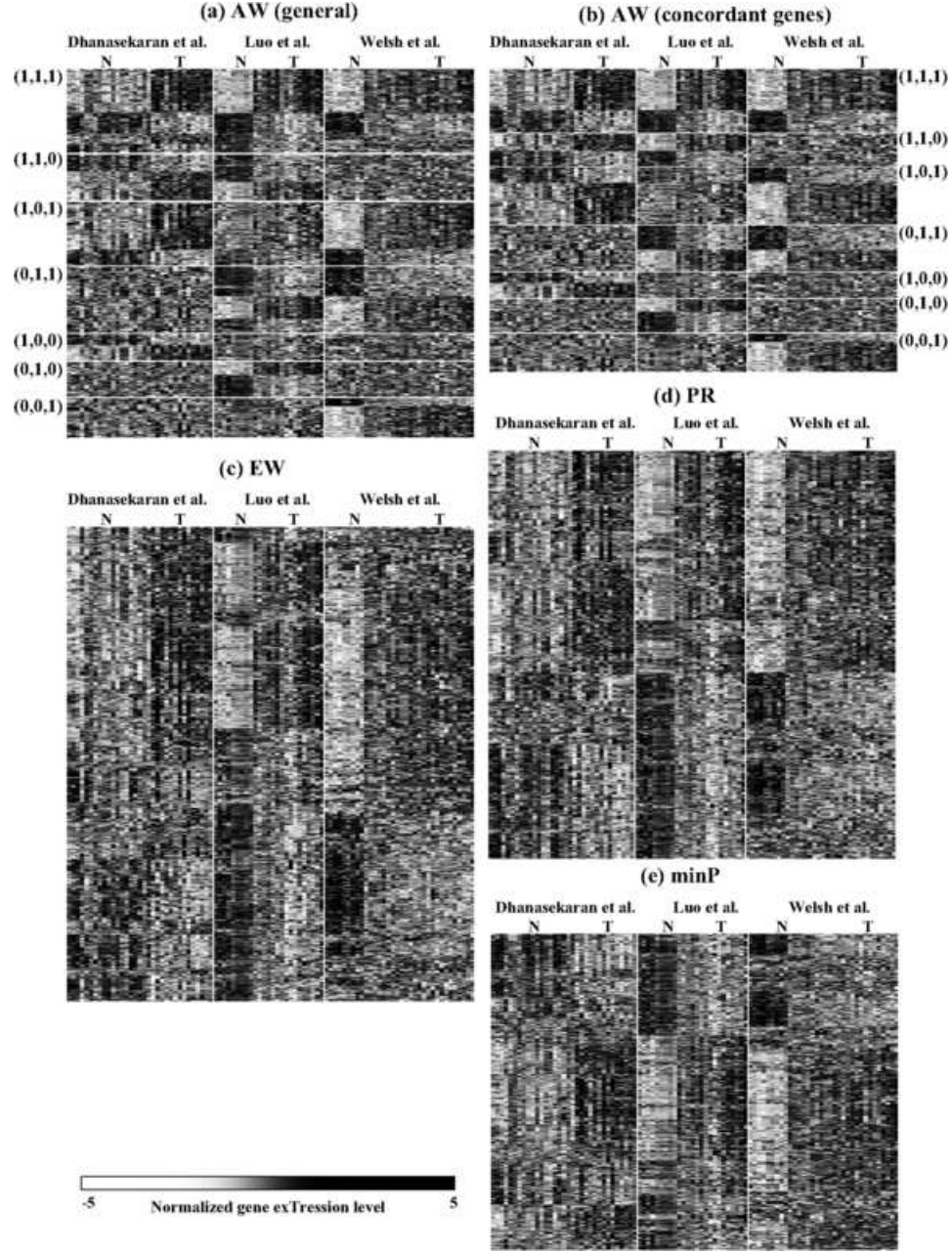


FIG. 2. Heatmaps of gene expression intensities for differentially expressed genes identified by different methods in the prostate cancer data sets.

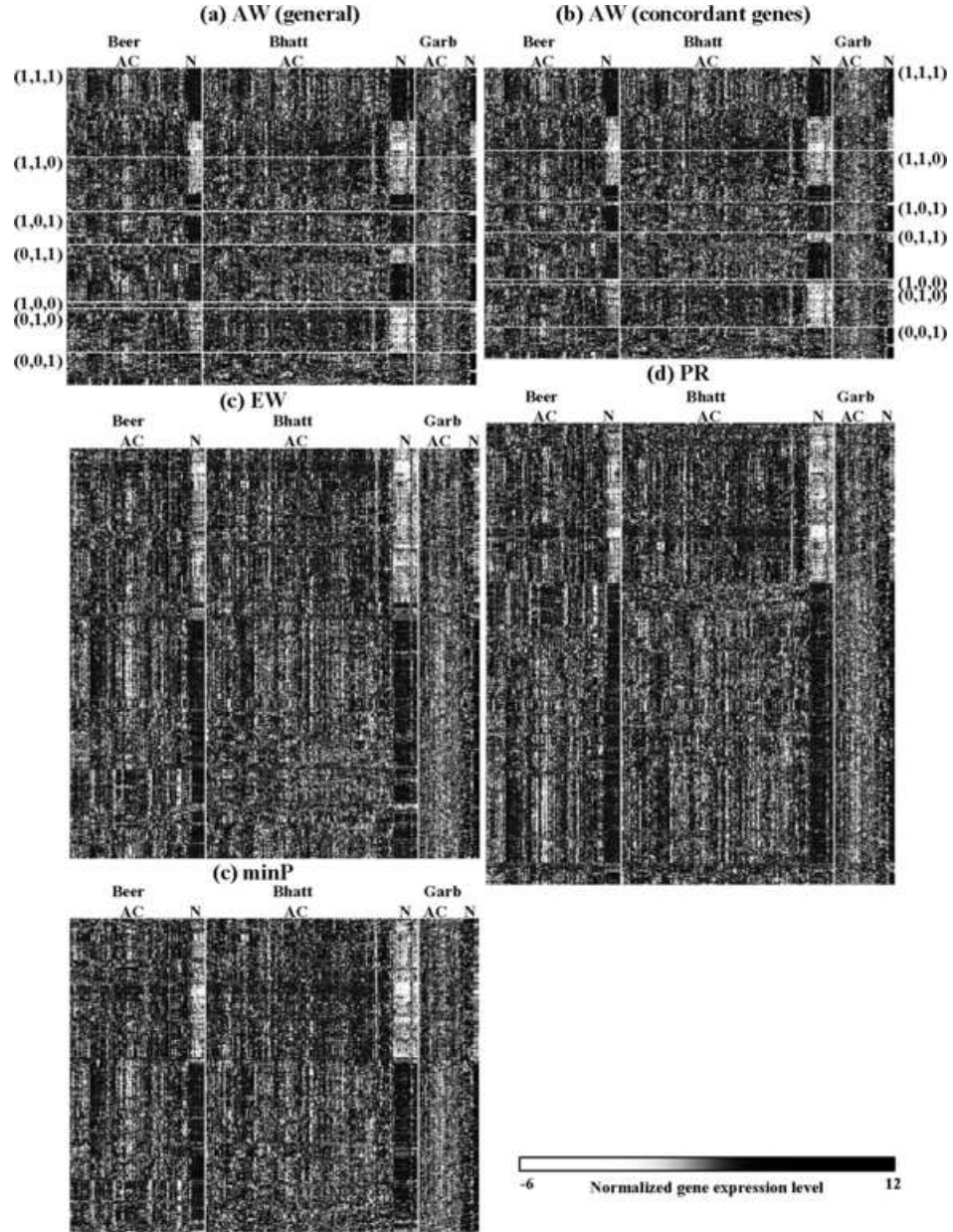


FIG. 3. Heatmaps of gene expression intensities for differentially expressed genes identified by different methods in the lung cancer data sets.

51 type (0, 1, 0) and 38 type (0, 0, 1). Overall, our lung cancer studies had more biomarkers that were consistent in terms of concordant up-regulation and down-regulation patterns, and fewer single study-specific biomarkers. These results match those from previous reports showing better consistency among lung cancer studies compared to prostate cancer studies, possibly due to larger sample sizes, better gene annotations, more specific disease subtype comparisons and better array quality. For example, Bhattacharjee and Beer used Affymetrix platforms, while Garber's data were generated from the lab of Pat Brown, the inventor of cDNA arrays.

**5. Power and admissibility.** In this section we drop the subscript  $g$  for genes and assume independence among studies when comparing five test statistics (EW, AW, minP, maxP and PR) for  $H_B$  at the univariate level. The maxP statistic is included for demonstration purposes although it is not targeted to  $H_B$ . To date, no best method for combining multiples studies has been identified, therefore, choosing a combined statistic must reflect specific biological purposes. Birnbaum (1954, 1955) established general conditions for evaluating combined methods, including monotonicity and admissibility. To compare several combined test procedures, he considered a one-sample test of the mean of a Gaussian distribution with known variance. We will use a similar two-sample test of the means of two Gaussian distributions with known variance:

$$(5.1) \quad Z_k = \frac{\bar{X}_{2k} - \bar{X}_{1k}}{\sigma_k \sqrt{1/n_{k1} + 1/n_{k2}}}, \quad k = 1, 2, \dots, K,$$

where  $\bar{X}_{1k} = (1/n_{k1}) \cdot \sum_{s=1}^{n_{k1}} X_{ks}$ ,  $\bar{X}_{2k} = (1/n_{k2}) \cdot \sum_{s=n_{k1}+1}^{n_{k1}+n_{k2}} X_{ks}$ ,  $X_{ks} \sim N(0, \sigma_k^2)$  when  $1 \leq s \leq n_{k1}$  and  $X_{ks} \sim N(\theta_k, \sigma_k^2)$  when  $n_{k1} + 1 \leq s \leq n_{k1} + n_{k2}$ . We will use the two-sided  $p$ -values  $P_k = \Pr(|Z| \geq |z_k| | \theta_k = 0)$  for study  $k$ , where  $Z$  is the standard normal distribution, to examine the acceptance regions of the various combined test procedures. The simplified framework is the focus for the discussion in the [Appendix](#) of admissibility and power comparisons of the five statistics. It is shown there that AW, EW, PR and minP are all admissible, but maxP is not.

5.1. *Power comparison of EW, AW, minP, maxP and PR under  $H_{Bh'}$ .* Denote by  $\Theta_0 = \{\theta_1 = \dots = \theta_K = 0\}$  and  $\Theta_A = \{\text{at least one } \theta_k \neq 0\}$  (i.e.,  $H_B$ ) the null and alternative hypothesis. Letting  $\beta^{\text{AW}}(\theta; \alpha)$  be the power of a test controlled at level  $\alpha$  for the OW statistic given  $\theta \in \Theta_A$ , we have

$$(5.2) \quad \beta^{\text{AW}}(\theta; \alpha) = \Pr(V^{\text{AW}} \leq C_\alpha^{\text{AW}} | \theta) = 1 - \int_{\Omega^{\text{AW}}} \prod_{k=1}^K p(P_k | \theta) dP_1 \cdots dP_K,$$

where  $C_\alpha^{\text{AW}}$  is the solution of  $v$  to the equation  $P(V^{\text{AW}} \leq v | \Theta_0) = \alpha$ ,  $\Omega^{\text{AW}} = \bigcap_{j=1}^{2^K-1} \{p(u(w_j)) > C_\alpha^{\text{AW}}\} = \bigcap_{j=1}^{2^K-1} \{U(w_j) < F_{\text{Gamma}(\sum_{k=1}^K w_{jk}, 1)}^{-1}(1 - C_\alpha^{\text{AW}})\}$  and  $F_{\text{Gamma}(\alpha, \beta)}^{-1}$  is the inverse CDF of a Gamma distribution with parameters  $\alpha$  and  $\beta$ ,  $w_j = (w_{j1}, \dots, w_{jK})$ ,  $w_{jk} \in \{0, 1\}$ ,  $k = 1, \dots, K$ , and enumeration index  $j$  exhausts all different weight vector possibilities such that  $\sum_{k=1}^K w_{jk} \geq 1$ . If the null hypothesis is true, it is generally accepted that the individual  $P_k$  value is uniformly distributed on  $[0, 1]$ . The density of the  $p$ -value under alternative law is expressed as

$$(5.3) \quad p(P|\theta) = \frac{p(x|\theta)}{p(x|0)} \Big|_{x=g(P)} \quad (0 \leq P \leq 1),$$

where  $x = g(P)$  indicates the solution of  $P = \int_x^1 f(x|0) dx$  [Pearson (1938)]. Similarly, the power for EW and minP can be calculated by  $\beta^{\text{EW}}(\theta; \alpha) = \int_{\Omega^{\text{EW}}} \prod_{k=1}^K p(P_k|\theta) dP_1 \cdots dP_K$ ,  $\beta^{\text{minP}}(\theta; \alpha) = 1 - [\int_{C_\alpha^{\text{minP}}}^1 p(P|\theta) dP]^K$  and  $\beta^{\text{maxP}}(\theta; \alpha) = [\int_{C_\alpha^{\text{maxP}}}^1 p(P|\theta) dP]^K$ , where  $\Omega^{\text{EW}} = \{-\sum_{k=1}^K \log p_k \geq C_\alpha^{\text{EW}}\}$ ,  $C_\alpha^{\text{EW}} = F_{\text{Gamma}(K, 1)}^{-1}(1 - \alpha)$ ,  $C_\alpha^{\text{minP}} = F_{\text{Beta}(1, K)}^{-1}(\alpha) = 1 - (1 - \alpha)^{1/K}$ ,  $C_\alpha^{\text{maxP}} = \alpha^{1/K}$ .

In our simplified setting, the  $Z$  test in (3) is used for power calculations, hence, the density of  $P_k$  is

$$(5.4) \quad \begin{aligned} p(P_k|\theta_k) &= \frac{1}{2} \exp\left\{\frac{c_k}{2}[2\Phi^{-1}(1 - P_k/2) - c_k]\right\} \\ &+ \frac{1}{2} \exp\left\{-\frac{c_k}{2}[2\Phi^{-1}(1 - P_k/2) + c_k]\right\}, \end{aligned}$$

where  $c_k = \frac{\theta_k}{\sigma_k \sqrt{1/n_{k1} + 1/n_{k2}}}$ ,  $k = 1, \dots, K$ . We consider  $n_{k1} = n_{k2} = 5$  and  $\sigma_k = 1$  so that the effect size is represented by  $\theta_k$  and power is evaluated with varying effect sizes.

The graphs in Figure 4 reflect a situation in which  $K = 10$  for simplified alternative hypothesis  $H_{Bh'}$  ( $1 \leq h \leq K$ ). Studies with nonzero effect sizes share a common effect size  $\theta$ . Power curves under  $\theta \in \{1.2, 1.4\}$  and varying values of  $h$  are displayed. Due to the difficulty of achieving an exact power calculation for  $K = 10$ , we performed 10,000 simulations to generate power curves. EW and AW are calculated for one-sided  $p$ -values for the purpose of comparability with PR, maxP, minP. In application, it is unlikely that the signs of effect will be known, therefore, two-sided  $p$ -values for maxP, minP, EW and AW are preferred. As expected, the figure shows that minP is more powerful than EW when  $h$  is small, and EW is more powerful than minP when  $h$  is large. On the other hand, AW performs stably and comparably to the best method in situations involving the two extremes. The performance



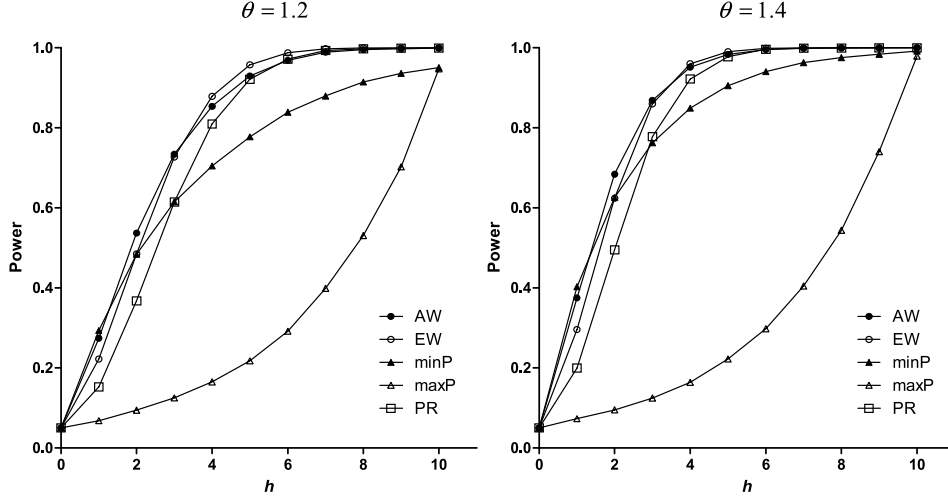


FIG. 4. Power analysis of EW, AW, minP, PR and maxP under  $H_{Bh'}$ ,  $1 \leq h \leq K$ . We compare power curves of the five methods combining  $K = 10$  studies. X axis represents  $h$ , the number of studies that have nonzero effects.

of maxP further confirms Loughin's conclusion that it has very low power unless  $h = K$ .

**6. Discussion.** In this paper we described our proposal for an adaptively weighted (AW) statistic for combining multiple studies, and reported our findings after applying it to two sets of combined microarray studies. Acknowledging that meta-analysis methods depend heavily on the biological question being investigated, we formulated two statistical hypothesis settings ( $HS_A$  and  $HS_B$ ) to identify differentially expressed genes considered significant in either partial or full data sets. Classical EW, minP and our proposed AW methods were used to analyze  $HS_A$ .

According to our findings, AW, EW and minP are all admissible in simplified scenarios. In terms of power analysis, EW was more powerful when all data sets were significant, while minP was more powerful when only one or a small number of data sets were significant. As a compromise between EW and minP, the AW method performed close to the best method in either extreme alternative hypothesis setting (Figure 5). Simulation results also confirmed this robust property of AW (Tables 3–5). In applications, AW had the additional advantage of categorizing differentially expressed genes by their adaptive weights, thus providing a practical basis for further biological exploration. In addition to not detecting discordantly regulated genes, the modified algorithm in Section 3, step IV, was appealing for the specific biological purpose of identifying all nondiscordant genes.

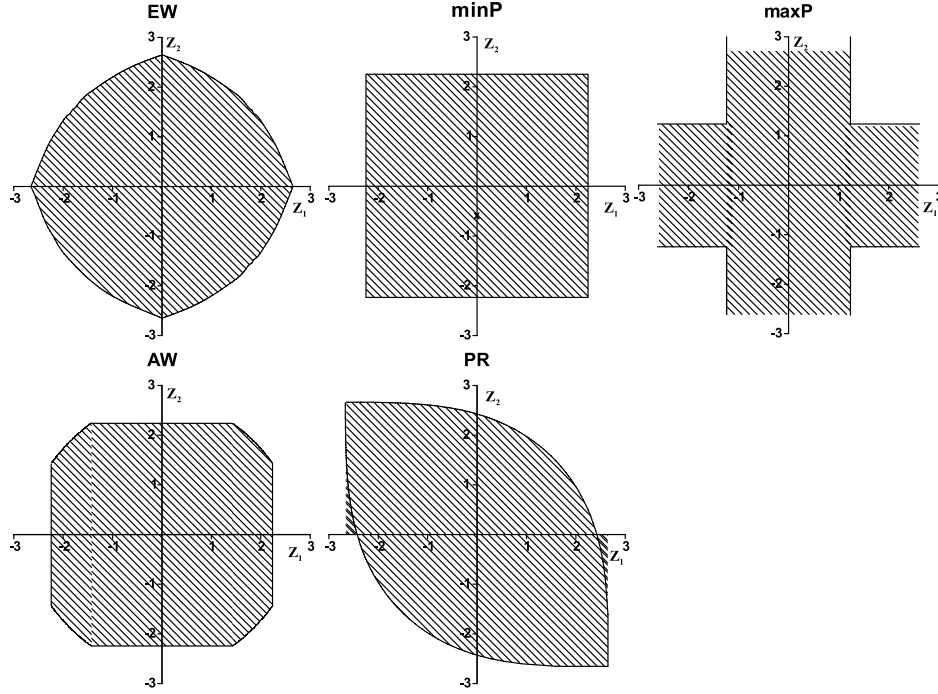


FIG. 5. Acceptance regions of EW, AW, minP, PR and maxP statistic for combining  $p$ -values from two independent studies when testing means of Gaussian distributions with known variances.

In this project we restricted the binary 0,1 adaptive weight search space for purposes of computational convenience and biological interpretability. For example, in Figure 1(b) the AW data support an immediate categorization of detected biomarkers, as well as information on similar/dissimilar differential gene expression between tissue pairs. As shown for the EW data in Figure 1(c), Fisher's method generated a large number of nontraceable biomarkers that were difficult to work with in terms of follow-up analyses. Theoretically, it is possible to extend the 0,1 space to a nonrestricted real number (i.e., positive weights that add up to 1). However, such results generate biomarker lists similar to those generated by the EW method [Figure 1(c)]. In other words, using nonbinary weights may be slightly superior statistically, but not biologically.

There are three limitations in addition to possible future extensions for future research. First, we assumed that all studies contain an identical matched gene list with no missing values. In actual practice, separate studies to be combined usually come from different microarray platforms. Requiring an identical matched gene list and no missing values will exclude many important genes that appear in certain studies but not in others, thus requiring

an extension that allows for missing values. Second, we focused on two-group comparison in this paper, and made a modification in order to limit detection to genes with concordant expression changes. To compare more than two groups, the  $F$ -statistic and its variations can be applied; resulting  $p$ -values from  $F$  tests can be combined similarly as described for the algorithm in Section 3. However, small  $p$ -values across studies do not guarantee concordant expression patterns. To address this problem, we have developed a multi-class correlation approach [Lu, Li and Tseng (2010)]. Third, our proposed method focuses on  $HS_B$  rather than  $HS_A$ , which is not the case with many biological applications. Finally, the AW statistic can be extended from biomarker detection to gene set enrichment analyses. Note that post-meta-analysis enriched pathways (gene sets) are thought to be more supportive of biological interpretations.

While we only considered combining multiple microarray studies in this paper, the methods we described can easily be extended to combinations of multiple genomic, epigenomic and/or proteomic studies—for instance, data sets from SNP arrays, genome arrays, methylation arrays, proteomic experiments and ChIP-on-chip experiments. Additional extensions and/or alternative models are required to accommodate biological knowledge and to address specific questions of interest.

## APPENDIX: ADMISSIBILITY

A test is considered admissible if it cannot be uniformly improved by any other test. No single test has been accepted as the most powerful, even in the simplified scenarios. Birnbaum expressed a necessary and sufficient condition (known as Theorem 5.1) for any test to be admissible under this situation.

**THEOREM 1** [Birnbaum (1954, 1955)]. *Under  $H_B$  and the test statistic is in the exponential family [e.g., equation (5.1)], the necessary and sufficient condition for a combined test procedure to be admissible is that the corresponding acceptance region is convex.*

Since the acceptance regions of EW and minP have been identified as convex, both methods are admissible; maxP is not. When proving that the PR method is admissible, Owen (2009) clarified Birnbaum’s (1954) misinterpretation of the PR method. The acceptance regions of EW, minP, maxP, AW and PR on the plane of a pair of  $Z$  statistics at level 0.05 are shown in Figure 5. When illustrating the rejection regions of several common combined tests (including EW and minP), Birnbaum showed a preference because it appeared to be “fairly sensitive in all directions.” From Figure 5, it is clear that the PR method prefers effects that show common directions in

two studies, since the rejection regions in the first and third quadrants are less stringent than the second and fourth quadrants. Note that AW actually shares positive aspects of both EW and minP methods: generally more sensitive than minP when parameters from both studies depart from 0 and more sensitive than EW when only one of the parameters departs from 0, and more sensitive than the minP method when parameters from both studies depart from 0. According to the following corollary, AW is admissible because the intersection of convex sets is convex, therefore, its acceptance region is convex.

**COROLLARY 1.** *The acceptance region of AW is convex and, thus, AW is admissible under  $H_B$  and assumption (5.1).*

**PROOF.** Denote by  $p_k = 2(1 - \Phi(|z_k|))$  the two-sided  $p$ -value, where  $\Phi(t) = \int_{-\infty}^t \phi(t) dt$ ,  $\phi(t)$  is the density of the standard normal distribution. First we prove that  $f(z_k) = -\log(p_k) = -\log(1 - \Phi(|z_k|)) + C$  is convex.  $f''(z) = \frac{\phi(|z|)}{[1 - \Phi(|z|)]^2} \{ \phi(|z|) - |z|[1 - \Phi(|z|)] \}$  when  $z \neq 0$ . It is well known that the elementary upper bound for  $1 - \Phi(x)$  is  $\phi(x)/x$ , for  $x > 0$ . Thus,  $f''(z) > 0$  when  $z \neq 0$ . Since  $f(z)$  is continuous at  $z = 0$ ,  $f(z)$  is convex in  $z$ . Hence,  $f(z_1, z_2, \dots, z_n) = -\sum_{k=1}^n \log(p_k)$  for any  $n \geq 1$  is convex, because the sum of convex functions is convex. For the AW statistic, the acceptance region is  $\{z_1, z_2, \dots, z_K : \min_{1 \leq k \leq K} p(u(w)) > c\}$ , where  $p(u(w))$  is the right-sided  $p$ -value of  $U(w)$ :

$$\begin{aligned} & \left\{ z_1, z_2, \dots, z_K : \min_{0 \leq k \leq K} p(u(w)) > c \right\} \\ &= \bigcap_{I_k \in \{0,1\}, 1 \leq k \leq K} \left\{ z_1, z_2, \dots, z_K : p \left( -\sum_{k=1}^K \log[p_k^{I_k}] \right) > c \right\} \\ &= \bigcap_{I_k \in \{0,1\}, 1 \leq k \leq K} \left\{ z_1, z_2, \dots, z_K : -\sum_{k=1}^K \log[p_k^{I_k}] < \gamma_j \right\}, \\ & \qquad \qquad \qquad j = 1, 2, \dots, 2^K - 1, \end{aligned}$$

$\gamma_j$  is  $F_{\text{Gamma}(\sum_{k=1}^K I_k, 1)}^{-1}(1 - c)$ . Thus, the acceptance region of AW is convex since the intersection of convex sets is also convex.  $\square$

**Acknowledgments.** The authors would like to thank Gerard Vockley for providing the mouse metabolism data set and reviewers for insightful comments.



## REFERENCES

- BEER, D. G., KARDIA, S. L., HUANG, C. C., GIORDANO, T. J., LEVIN, A. M., MISEK, D. E., LIN, L., CHEN, G., GHARIB, T. G., THOMAS, D. G., LIZYNESS, M. L., KUICK, R., HAYASAKA, S., TAYLOR, J. M. G., IANNETTONI, M. D., ORRINGER, M. B. and HANASH, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.* **8** 816–824.
- BERGER, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24** 295–300. [MR0687187](#)
- BHATTACHARJEE, A., RICHARDS, W. G., STAUNTON, J., LI, C., MONTI, S., VASA, P., LADD, C., BEHESHTI, J., BUENO, R., GILLETTE, M., LODA, M., WEBER, G., MARK, E. J., LANDER, E. S., WONG, W., JOHNSON, B. E., GOLUB, T. R., SUGARBAKER, D. J. and MEYERSON, M. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* **98** 13790–13795.
- BIRNBAUM, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49** 559–574. [MR0065101](#)
- BIRNBAUM, A. (1955). Characterizations of complete classes of tests of some multiparametric hypotheses, with applications to likelihood ratio tests. *Ann. Math. Statist.* **26** 21–36. [MR0067438](#)
- BOROVECKI, F., LOVRECIC, L., ZHOU, J., JEONG, H., THEN, F., ROSAS, H. D., HERSCHE, S. M., HOGARTH, P., BOUZOU, B., JENSEN, R. V. and KRAINC D. (2005). Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc. Natl. Acad. Sci. USA* **102** 11023–11028.
- CARDOSO, J., BOER, J. H., MORREAU, H. and FODDE, R. (2007). Expression and genomic profiling of colorectal cancer. *Biochim. Biophys. Acta Rev. Cancer* **1775** 103–137.
- CHOI, H., SHEN, R., CHINNAIYAN, A. M. and GHOSH, D. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* **8** 364–383.
- CHOI, J. K., YU, U., KIM, S. and YOO, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19** 84–90.
- CONLON, E. M., SONG, J. J. and LIU, A. (2007). Bayesian meta-analysis models for microarray data: A comparative study. *BMC Bioinformatics* **8** 80–100.
- CONLON, E. M., SONG, J. J. and LIU, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics* **7** 247–250.
- COUSINS, R. D. (2007). Annotated bibliography of some papers on combining significances or  $p$ -values. Available at [arXiv:0705.2209v1](#).
- DAI, M., WANG, P., BOYD, A. D., KOSTOV, G., ATHEY, B., JONES, E. G., BUNNEY, W. E., MYERS, R. M., SPEED, T. P., AKIL, H., WATSON, S. J. and MENG, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res.* **33** e175. doi: [10.1093/nar/gni179](#).
- DHANASEKARAN, S. M., BARRETTE, T. R., GHOSH, D., SHAH, R., VARAMBALLY, S., KURACHI, K., PIENTA, K. J., RUBIN, M. A. and CHINNAIYAN, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412** 822–826.
- EFRON, B., TIBSHIRANI, J. D., STOREY, R. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. [MR1946571](#)
- FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4 ed. Oliver and Boyd, Edinburgh.
- GARBER, M. E., TROYANSKAYA, O. G., SCHLUENS, K., PETERSEN, S., THAESLER, Z., PACYNA-GENGELBACH, M., VAN DE RIJN, M., ROSEN, G. D., PEROU, C. M., WHYTE, R. I., ALTMAN, R. B., BROWN, P. O., BOTSTEIN, D. and PETERSEN, I.

- (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA* **98** 13784–13789.
- GEORGE, E. O. (1977). Combining independent one-sided and two-sided statistical tests—some theory and applications. Ph.D. thesis, Univ. Rocheser. [MR2627130](#)
- GHOSH, D., BARRETTE, T. R., RHODES, D. and CHINNAIYAN, A. M. (2003). Statistical issues and methods for meta-analysis of microarray data: A case study in prostate cancer. *Functional and Integrative Genomic* **3** 180–188.
- GOOD, I. J. (1955). On the weighted combination of significance tests. *J. Roy. Statist. Soc. Ser. B* **17** 264–265. [MR0076252](#)
- HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, New York. [MR0798597](#)
- HONG, F., BREITLING, R., MCENTEE, C. W., WITTNER, B. S., NEMHAUSER, J. L. and CHORY, J. (2006). Rankprod: A bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22** 2825–2827.
- HU, P., GREENWOOD, C. M. T. and BEYENE, J. (2005). Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics* **6** 128–138.
- LANCASTER, H. (1961). The combination of probabilities: An application of orthonormal functions. *Austr. J. Statist.* **3** 20–33. [MR0130742](#)
- LITTELL, R. C. and FOLKS, J. L. (1971). Asymptotic optimality of Fisher’s method of combining independent tests. *J. Amer. Statist. Assoc.* **66** 802–806. [MR0312634](#)
- LITTELL, R. C. and FOLKS, J. L. (1973). Asymptotic optimality of Fisher’s method of combining independent tests, ii. *J. Amer. Statist. Assoc.* **68** 193–194. [MR0375577](#)
- LOUGHIN, T. M. (2004). A systematic comparison of methods for combining  $p$ -values from independent tests. *Comput. Statist. Data Anal.* **47** 467–485. [MR2086483](#)
- LU, S., LI, J., SONG, C., SHEN, K. and TSENG, G. C. (2010). Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics* **26** 333–340.
- LUO, J., DUGGAN, D. J., CHEN, Y., SAUVAGEOT, J., EWING, C. M., BITTNER, M. L., TRENT, J. M. and ISAACS, W. B. (2001). Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Res.* **61** 4683–4688.
- MOREAU, Y., AERTS, S., DE MOOR, B., DE STROOPER, B. and DABROWSKI, M. (2003). Comparison and meta-analysis of microarray data: From the bench to the computer desk. *Trends Genet.* **19** 570–577.
- OLKIN, I. and SANER, H. (2001). Approximations for trimmed Fisher procedures in research synthesis. *Statist. Methods Med. Res.* **10** 267–276.
- OWEN, A. B. (2009). Karl Pearson’s meta-analysis revisited. *Ann. Statist.* **37** 3867–3892. [MR2572446](#)
- PARMIGIANI, G., GARRETT-MAYER, E. S., ANBAZHAGAN, R. and GABRIELSON, E. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin. Cancer Res.* **10** 2922–2927.
- PEARSON, E. S. (1938). The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* **30** 134–148.
- PEARSON, K. (1934). On a new method of determining ‘goodness of fit.’ *Biometrika* **26** 425–442.
- PIROOZNIA, M., NAGARAJAN, V. and DENG, Y. (2007). Gene venn—a web application for comparing gene lists using venn diagram. *Binformatics* **1** 420–422.
- RHODES, D., BARRETTE, T. R., RUBIN, M. A., GHOSH, D. and CHINNAIYAN, A. M. (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* **62** 4427–4433.

- ROY, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24** 220–238. [MR0057519](#)
- SEGAL, E., FRIEDMAN, N., KOLLER, D. and REGEV, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genet.* **3** 1090–1098.
- SHEN, R., GHOSH, D. and CHINNAIYAN, A. M. (2004). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* **5** 94–109.
- SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.* **3** Article 3. [MR2101454](#)
- STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* **64** 479–495.
- STOUFFER, S., SUCHMAN, E., DEVINNERY, L., STAR, S. and WILLIAMS, J. (1949). *The American Soldier, Vol. I: Adjustment during Army Life*. Princeton Univ. Press, Princeton, NJ.
- TIPPETT, L. H. C. (1931). *The Methods in Statistics*, 1st ed. Williams and Norgate, London.
- TSENG, G. C., OH, M. K., ROHLIN, L., LIAO, J. C. and WONG, W. H. (2001). Issues in cdna microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29** 2549–2557.
- TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** 5116–5121.
- WELSH, J. B., SAPINOSO, L. M., SU, A. I., KERN, S. G., WANG-RODRIGUEZ, J., MOSKALUK, C. A., FRIERSON, H. F. and HAMPTON JR., G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.* **61** 5974–5978.
- WILKINSON, B. (1951). A statistical consideration in psychological research. *Psychol. Bull.* **48** 156–157.

DEPARTMENT OF BIOSTATISTICS  
UNIVERSITY OF PITTSBURGH  
PITTSBURGH, PENNSYLVANIA  
USA  
E-MAIL: [jli3@hfh.org](mailto:jli3@hfh.org)  
[ctseng@pitt.edu](mailto:ctseng@pitt.edu)